FR-04-61

HumRRO

Human Resources Research Organization

Third-Party Checking of 2004 Scaling and Equating for the Kentucky Core Content Test

Arthur A. Thacker
Andrea Sinclair
R. Gene Hoffman

Human Resources Research Organization (HumRRO)
950 Breckenridge Lane, Suite 170
Louisville, KY 40207
Phone (502) 721-9045
FAX (502) 721-9983

Prepared for:

Kentucky Department of Education
Capital Plaza Tower, 18th Floor
500 Mero Street
Frankfort, KY 40501

September 2004

Third-Party Checking of 2004 Scaling and Equating for the Kentucky Core Content Test

Table of Contents

Summary

CTB and HumRRO independently calculated the scaled and equated raw-score-to-scale-score tables for the 2004 Kentucky Core Content Test (KCCT). From those tables, cut points were identified that can be used to assign student performance classifications (Novice, Apprentice, Proficient, or Distinguished (NAPD)). Student scores can then be aggregated to the school level to determine each school's accountability index.

In addition, the 2004 and subsequent administrations of the KCCT will be used to determine if schools have met the Adequate Yearly Progress (AYP) requirements of the federal No Child Left Behind (NCLB) Act. Because of the reporting timeline and other requirements of NCLB, the 2004 KCCT required two rounds of scaling and equating. The first round assigned students' scores base only on their performance on KCCT's machine scored (multiple-choice) items. This allowed for a more rapid reporting schedule resulting in preliminary AYP determinations. The calculation of preliminary results also required separate scaling and equating procedures for reading and math components of the KCCT. Raw-score-to-scale-score tables were created for the multiple-choice-only version of KCCT as well as for the full assessment.

An attempt was made to use an early-return sample of school districts for scaling and equating purposes for the full assessment. However, due to issues surrounding item parameter estimation and sample selection, the full sample of students was used instead. Previously established rules regarding students' inclusion in the calibration sample were followed for 2004.

Decisions regarding the handling of problem test items were discussed between CTB and HumRRO and in all cases both groups reached consensus. Results calculated by HumRRO were nearly identical to those calculated by CTB. Given that our scaling and equating results were nearly identical (small differences due to rounding, etc., that would not affect any students NAPD classifications) with those of CTB, we are assured that CTB did not commit processing errors.

Third-Party Checking of 2004 Scaling and Equating for the Kentucky Core Content Test

Introduction

Every year, the Kentucky Core Content Test (KCCT) [1] is scaled and equated by Item Response Theory (IRT) using a calibration sample of students in designated grades (4, 5, 7, 8, 10, and 11). Scaling involves the estimation of item parameters for the current year's test. These item parameters are linearly transformed to a 325-800 point scale and equated with previous years' scales. The results of scaling and equating are then used to construct raw-score-to-scale-score tables for every KCCT test form. Cut points are also identified so that students' raw scores can be translated to performance categories: Novice, Apprentice, Proficient, and Distinguished (NAPD).

Scaling and equating are done for the following grade/subject combinations:

Grade 4 - Reading, Science
Grade 5 - Math, Social Studies, Arts & Humanities, Practical Living/Vocational Studies
Grade 7 - Reading, Science
Grade 8 - Math, Social Studies, Arts & Humanities, Practical Living/Vocational Studies
Grade 10 - Reading, Practical Living/Vocational Studies
Grade 11 - Math, Science, Social Studies, Arts & Humanities

As a quality control step, personnel at CTB and the Human Resources Research Organization (HumRRO) conduct scaling and equating analyses simultaneously and independently. Researchers at both companies compare results at several steps throughout the process. If a result between CTB and HumRRO is not identical, then procedures are reviewed until the issue is resolved and both staffs get the same outcome. This way, the complex sampling, item parameter estimation analyses, Stocking-Lord equating, raw-score-to-scale-score transformations, and cut point identifications are checked and verified by two autonomous agencies.

The procedures used by HumRRO are outlined in detail below.

[1] The test in use before 1998 was the Kentucky Instructional Results Information System (KIRIS) test.

Creation of a Multiple-Choice-Only Scale

A multiple-choice-only scale was used in 2004 as a one-time-only solution to make preliminary AYP decisions. This scale was used to meet reporting deadlines for NCLB. Future administrations of KCCT will use all items for this purpose. In order to generate scores for students using only multiple-choice items, a special scale was required. To link that scale with both the full set of items and with previously administered KCCT tests, a special multiple-choice-only scale was created using data from 2002. Briefly, the 2002 results for multiple-choice items were used to generate item parameters independently of the open-response items. Those item parameters were then equated via a Stocking/Lord procedure to the 2002 operational scale. The transformed parameters were used as anchor items to link the 2004 multiple-choice-only results to the 2002 operational scale. Full details are provided in a previous report (Thacker, 2004).

Sample Identification and File Construction

The first step in performing the required analyses was to identify a calibration sample for each grade/subject and construct files formatted for use with CTB's IRT programs (Pardux and Flux). This process was necessary for both the multiple-choice-only and for the full-item-set data. The procedures used to read CTB's Winscore (machine-scored data) were established in 2003 and altered to accommodate changes in the file's structure in 2004. Other changes were required to create readable files for IRT processing of the multiple-choice-only data.

In past years, Kentucky has selected most of its student population for use in the calibration sample for scaling and equating. However, some students are purposefully exempted (a student who leaves the test form completely blank, for example). CTB has devised a set of rules for including students in the calibration file based on KDE's recommendations and the CTB file structure. HumRRO independently wrote a SAS program to apply those rules. An attempt was made this year to use an early-return sample of school districts for scaling and equating purposes for the full-item set. The purpose of the early-return schools was to allow calibration and equating to happen earlier in the calendar year, and thereby report scores earlier as well. However, due to issues surrounding item parameter estimation and sample selection, the full sample of students was used instead. Previously established rules regarding students' inclusion in the calibration sample were followed for 2004. CTB and HumRRO compared results at several stages during this procedure and most differences in the two sets of files were resolved. However, in one case, CTB's and HumRRO's calibration samples were consistently different by a single student. This student had an incomplete record for hand-scored items. HumRRO assigned a score of 0 to the missing items, while CTB assigned a blank. There is no convention for dealing with students with partial records, and this issue has not been encountered previously. Records for students should be populated for all test items, even if the student does not attempt particular items. CTB agreed to investigate after the calibration and scaling were complete. Due to the short time allotted for this procedure and the relatively large number of students included in Kentucky's calibration sample (> 40,000 for all grade/subjects) CTB and HumRRO decided to eliminate the anomalous student from the calibration files. HumRRO and CTB verified that the samples were identical prior to beginning IRT processing.

A second anomaly in the data was discovered during sample identification. In 2004, the entire data set was posted as two files, one for hand-scored and one for machine-scored data. The machine-scored (Winscore) file was used for both the multiple-choice-only and full-item-set processing. HumRRO and CTB used different conventions initially to select calibration sample students. In previous years, students were placed in the calibration sample if they had responded to at least one multiple-choice item. They were not required to correctly answer the one item, only to have attempted it. CTB began with a requirement that students receive at least one correct response in 2004 while HumRRO adhered to the previous rule. It was decided by consensus between CTB, KDE, and HumRRO to revert to the previous rule to keep equating consistent. In addition, changes in the variables included in the Winscore files led to early differences in the sample selected by CTB and HumRRO. These two issues necessitated restarting the process several times by both CTB and HumRRO.

A third anomaly that occurred during this process was an apparent disparity in the numbers of students selected for the calibration sample. CTB's original files for each grade/subject contained about 50-70 fewer students than HumRRO's files. This difference was attributed to a program error, whereby CTB's files were being truncated and the last few records deleted. CTB repaired the error and its sample sizes matched HumRRO's.

A fourth issue, of which both CTB and HumRRO were aware prior to receiving student data files, was that some multiple-choice items were coded differently than others. In 2004 students marked their answers directly in their test booklets. In the past, there were corresponding answer sheets with letters or numbers for each choice in the test booklet. The letters/numbers were omitted from the test booklets in 2004 and students marked the circles near their choices. There was some confusion between CTB and a subcontractor regarding the creation of the key documents for the test booklets. Responses are typically arranged in a box pattern (2 answers per row, 2 rows per item, for 4 potential choices). The confusion occurred when the keys were made under either the assumption that the first column of answers represented A, and B, or whether the first row of answers represented A and B. For some items the B and C responses were switched in the key documents. CTB and their subcontractor provided a variable that indicated whether items followed this N or Z pattern in the key documents. HumRRO wrote special lines of code to detect this variable and make the necessary changes to score all items correctly.

Finally, when CTB and HumRRO were both operating on the assumption that only the early return districts would be used for calibration, it was necessary to select students based only from those districts. Early comparisons between samples showed that CTB often had 1-14 students more in this file than HumRRO. An error in HumRRO's SAS program that read the district code caused this problem. This error became a non-issue once the decision was made to include the full population rather than the early return sample. However, the error in the program was corrected in order to facilitate an early return sample strategy in subsequent years.

Scaling and Equating Procedures

Item response data for all of the 2004 test forms were scaled using CTB's PARDUX program. Based on IRT, PARDUX uses a three-parameter logistic model for multiple-choice

items and a two-parameter model for open-response items to estimate item parameters. Item parameters from both these models are eventually transformed to a single scale.

The equating process involves the application of the Stocking-Lord procedure to two different sets of anchor[2] item parameters: anchor item parameters from 2002 and anchor item parameters from 2004. These two sets of parameters are on different metrics. The 2004 parameters are on a theta metric (-1 to +1 scale) and the 2002 item parameters are on the "Kentucky metric" (325 to 800 scale). Stocking-Lord produces transformation constants (M1 and M2) that are used to linearly transform the 2004 metric onto the 2002 metric. This transforms all the 2004 item parameters onto the 325 – 800 scale, which traces back to the original 1992 scale.

One other issue should be mentioned with regard to the production of anchor files. CTB uses the FLUX program and several "hand-steps" in order to create anchor files. HumRRO uses a SAS program written specifically to produce the anchor files. CTB's and HumRRO's anchor files contained slight differences in the last decimal place for several parameters. Investigation of the differences revealed that Flux truncates at the last decimal while HumRRO's SAS program rounds. HumRRO discovered this anomaly in 2002 and created anchor files with Flux and with the SAS program to investigate whether there were meaningful differences caused by these slight inconsistencies. There were none in 2002. In 2003 and 2004, HumRRO used only the anchors created using the SAS program. Again, the slight differences in M1, M2, and in student scoring tables would have caused no differences in student or school classifications.

The final step in the process is to use CTB's FLUX program to create raw-score-to-scale-score conversion tables and identify the cut points for the performance categories. The slight variations in M1 and M2 did cause some small differences in CTB's and HumRRO's scoring tables, but never by more than one scale score point and in no instance did the differences affect student performance classifications. An anomaly discovered during this phase was that CTB had inadvertently used the smaller sample (less the 50-70 students) as described above for processing Grade 10 reading. Because of the late discovery of this issue, CTB decided to use that sample as the operational solution and requested that HumRRO duplicate those results. However, as HumRRO attempted to duplicate CTB's process, another issue was discovered. CTB had used the transformation constants from the full sample to create the raw-score-to-scale-score tables, but had used parameters estimated from the truncated sample. When HumRRO discovered this error, CTB agreed to correct the raw-score-to-scale score table for Grade 10 reading using the full sample of students. CTB's final solution for Grade 10 reading matched HumRRO's original solution exactly.

In 2004, because of the multiple-choice-only scale, two sets of anchors were used for reading and mathematics. Both sets of anchors come from the 2002 data, but one set was created specifically for the multiple-choice-only scale while the other was created from operational 2002 item parameters. This allowed CTB and HumRRO to link both versions of the results from 2004 to 2002 and to the original 1992 scale.

[2] Anchor items were designated on one form for each grade/subject on the 2001 KCCTs. The same anchor form was readministered in 2003 with all items intact and occurring in the same sequence as in 2001.

Scope of Third-Party Checking

In addition to doing a parallel analysis with CTB this year, HumRRO also conducted an in-house, parallel analysis to accomplish scaling and linking for the 2004 data. The Processing Steps listed below, while adequate, are being improved each year to ensure greater accuracy, standardization, and efficiency. This year, because of the changes in the student data files HumRRO received, a large portion of HumRRO's efforts were dedicated to reading the new files and creating calibration files correctly formatted for use in IRT processing programs. The multiple-choice-only scaling processes, which were implemented on a one-time-only basis, also increased the amount of effort required to ensure accurate data processing and student scoring.

Processing Steps

HumRRO took the following steps for each grade/subject tested (all listed processes were required for both the multiple-choice-only and full-item-set procedures):

1. Created anchor files (PARDUX *.anc) of multiple-choice test items that appeared on the anchor form. These anchor items were used to equate the 2003 test to the 2001 scale. The 2004 anchor files were created using 2002 parameter files for the matching forms.

2. Created working files (PARDUX *.RWO) from the calibration sample for the 2003 Kentucky Core Content Test. These files include both open-response and multiple-choice data.

3. Prepared control files (PARDUX *.ctl) which contain the constraints used for item parameter estimation, student proficiency estimation, maximum number of items, etc. The SAS program used to create *.rwo files included a routine to print out a control file.

4. Estimated parameters for Kentucky Core Content Test items using PARDUX.

5. Performed Stocking-Lord transformation using PARDUX. The results of this transformation include a slope and intercept constant for equating the 2004 Kentucky Core Content Test back to 2002.

6. Confirmed that the equating constants (M1 and M2) from Step 5 matched those derived by CTB.

7. Created parameter files (FLUX *.par) for each test form for use in preparation of raw-score-to-scale-score tables. A special SAS program was written for this purpose.

8. Created files (FLUX *.hlk) containing the scale limits (325 and 800) and constants from the Stocking-Lord transformation. A special SAS program was written for this purpose.

9. Created raw-score-to-scale-score transformation tables for each form using FLUX.

10. Confirmed that the raw-score-to-scale-score transformation tables from Step 9 match those derived by CTB and verified cut points used to separate student performance into Novice (Non-performing, Middle, High)/Apprentice (Low, Middle, High)/Proficient/Distinguished categories.

Results

After performing periodic checks with CTB as individual tests were scaled and equated, HumRRO and CTB reached near-exact agreement on the equating constants for all grade/subjects. Table 1 summarizes the results of this study for the multiple-choice-only scale. Schools' preliminary AYP determinations are made using these results. Table 2 summarizes the results of this study for the full item set. Schools' final AYP determinations and accountability indexes are made using these results. Tables 1 and 2 follow the same format. Grade and subject are identified for each test in the first two columns, respectively. The stage at which convergence occurred (if at all) is recorded in the third column. If convergence was not reached after 50 iterations by the PARDUX program, the solution at Stage 50 was accepted by mutual agreement. The fourth column identifies problem items and references the solutions that were reached cooperatively between CTB and HumRRO. The next four columns contain the M1 and M2 (slope and intercept) constants obtained from the Stocking-Lord transformation. CTB computed the first set of constants and HumRRO the second. The ninth column contains the difference between CTB's and HumRRO's M1 constants (i.e., $M1_{CTB}$-$M1_{HumRRO}$). The tenth column records the same information for M2 constants (i.e., $M2_{CTB}$-$M2_{HumRRO}$).

The last two columns in Tables 1 and 2 list whether there was exact agreement between CTB and HumRRO on (1) the raw-score-to-scale-score tables and (2) the cut points. Cut points from these tables are used to assign students to performance categories that, in turn, are used in the computation of each school's accountability index and AYP determination. CTB and HumRRO were in near-exact agreement for all raw-score-to-scale-score tables for every grade/subject for both scaling procedures.

Explanations of convergence issues and individual item issues are footnoted in Table 2. The footnotes explain the specific problems and their solutions when there isn't sufficient space in the tables. It should be noted that all problem items were dealt with during the parameter estimation phase of the scaling and equating process. No anchor item for which parameters were estimated was eliminated from the Stocking-Lord procedure. Item 146 was removed from the Grade 11 mathematics test for both scaling procedures. This multiple-choice item appeared only on Forms 6A and 6B. Scoring tables were adjusted accordingly.

Table 1. KCCT 2004 Results: Multiple-Choice Items Only

				CTB		HUMRRO		CTB-HUMRRO Differences			
Grade	*Subject*	*Convergence*	*Problems*	*M1*	*M2*	*M1*	*M2*	*M1*	*M2*	*Score Tables Agreement*	*NAPD Exact Agreement*
4	RD	Stage 18	None	31.92660	553.93121	31.92649	553.93103	0.00011	0.00018	yes	yes
7		Reached max	Convergence	30.41715	515.97003	30.41743	515.97021	-0.00028	-0.00018	yes	yes
10		Stage 19	None	55.01707	514.83636	55.01703	514.83636	0.00004	0.00000	yes	yes
5	MA	Stage 20	None	34.41101	574.79541	34.41096	574.79541	0.00005	0.00000	yes	yes
8		Reached max	Convergence	30.85264	540.87518	30.85269	540.87518	-0.00005	0.00000	yes	yes
11		Stage 24	Item 146, decision to remove this item	41.22234	544.10376	41.22232	544.10376	0.00002	0.00000	yes	yes

Table 2. KCCT 2004 Results: Full-Item Set (MC +OR Items)

				CTB		HUMRRO		CTB-HUMRRO Differences			
Grade	*Subject*	*Convergence*	*Problems*	*M1*	*M2*	*M1*	*M2*	*M1*	*M2*	*Score Tables Agreement*	*NAPD Exact Agreement*
4	RD	Stage 14	None	30.26125	554.56482	30.26129	554.56470	-0.00004	0.00012	Form1 rs=66; 1 ss pt diff	yes
	SC	Stage 15	None	25.82720	557.26721	25.82717	557.26721	0.00003	0.00000	yes	yes
5	A&H	Stage 15	None	43.83070	534.06348	43.83090	534.06360	-0.00020	-0.00012	yes	yes
	MA	Stage 15	None	34.88600	574.76910	34.88519	574.76648	0.00081	0.00262	Form1 rs=39; 1 SE pt diff	yes
	PL	None	Convergence	45.30069	521.81415	45.30091	521.8139	-0.00022	0.00025	yes	yes
	SS	Stage 14	None	31.48694	551.73328	31.48688	551.73322	0.00006	0.00006	yes	yes
7	RD	Stage 16	Item 150, Extra M-Step	26.79676	517.83539	26.79679	517.83521	-0.00003	0.00018	yes	yes

Table 2. KCCT 2004 Results: Full-Item Set (MC +OR Items)

				CTB		HUMRRO		CTB-HUMRRO Differences			
Grade	*Subject*	*Convergence*	*Problems*	*M1*	*M2*	*M1*	*M2*	*M1*	*M2*	*Score Tables Agreement*	*NAPD Exact Agreement*
	SC	Stage 19	None	26.53054	510.19711	26.53064	510.19705	-0.00010	0.00006	yes	yes
8	A&H	Stage 19	None	49.42603	521.35498	49.42598	521.35498	0.00005	0.00000	yes	yes
	MA	Stage 30	None	31.21453	539.74805	31.21450	539.74805	0.00003	0.00000	yes	yes
	PL	Stage 15	None	38.49185	506.16681	38.49168	506.16699	0.00017	-0.00018	yes	yes
	SS	Stage 15	None	39.84921	522.65961	39.84937	522.65906	-0.00016	0.00055	Form1 rs=6; 1 ss pt diff	yes
10	PL	Stage 35	Item 90,Extra M-Step	45.81382	512.64624	45.81382	512.64612	0.00000	0.00012	yes	yes
	RD	None	Items 5, 35, 40, 45, 105, 115 and 120, Extra M-Step	47.28650	512.79724	47.28651	512.79724	-0.00001	0.00000	yes	yes
11	A&H	None	Convergence and Item 90,Extra M-Step	58.60991	538.78192	58.60991	538.78204	0.00000	-0.00012	yes	yes
	MA	Stage 29	Item 146 removed and Item 97,Extra M-Step	40.71188	539.1037	40.7118	539.1037	0.00008	0.00000	yes	yes
	SC	Stage 18	None	32.04132	545.68341	32.04129	545.68341	0.00003	0.00000	yes	yes
	SS	Stage 19	Item 105, Extra M-Step	50.82182	547.31061	50.82172	547.31061	0.00010	0.00000	yes	yes

HumRRO also verified the cut points on the raw-score-to-scale-score tables. Cut points were assigned by rule. HumRRO verified cut points between Novice and Apprentice, between Apprentice and Proficient, and between Proficient and Distinguished performance categories. HumRRO also verified cut points for Low, Medium, and High subcategories within the Novice and Apprentice categories. In one instance HumRRO and CTB assigned different performance categories when both the raw score and scale score were identical. In this instance, HumRRO had assigned students a category of Novice-Medium and CTB had assigned a category of Novice-Low (or Non-performing). By rule, only students scoring near chance are assigned this lowest of categories (they receive the minimum scale score of 325). The scale score in question was just slightly more than that minimum. CTB agreed to reassign that scale score a category of Novice-Medium.

Documentation

To document the steps involved in scaling and linking the 2004 Kentucky Core Content Test, HumRRO saved all electronic files used in data preparation, including SAS programs, SAS logs, and SAS output lists and all files produced during PARDUX scaling and FLUX transformations. These files have been submitted to the Kentucky Department of Education (KDE). Appendices from the Hoffman and Thacker (1999) report contain hardcopy examples of important files that were submitted.

Files were submitted for both the MC only and full-item-set processing. The files were submitted on two separate CDs in order to limit the potential for mistaking one set of files from the other. All MC only files are located in directories labeled as such on the CD. Naming conventions for the output files are identical due to the automated steps involved in processing.

All electronic files submitted to KDE are named according to the following code (where S = subject, G = grade level).

A. PARDUX Control File (SSGG04.CTL). This file contains the number of items, the maximum number of stages for PARDUX, the convergence criterion, parameter estimation limits, and maximum and minimum values for proficiency estimates (theta). It also contains information allowing the program to distinguish between open-response and multiple-choice items, the number of score levels for open-response data, and which items to include in parameter estimation.

B. PARDUX Data File (SSGG04.RWO). This file contains the student score data. It is coded such that a 1 indicates a correct answer for a multiple-choice question and actual score levels (0-4) are recorded for student responses to open-response questions. To facilitate communication, HumRRO adhered to CTB's item order in constructing these data files.

C. PARDUX Anchor File (SSGG04.ANC). This file contains common-scaling item parameters from the 2001 KCCT (the identical items appeared on the 2002 KCCT). Only multiple-choice items are used in *.ANC files.

D. SAS Programs configured as SSGGrwcd.sas. This program produces the anchor files (*.ANC), PARDUX control files (*.CTL), and student score files (*.RWO). The SAS log and list files generated by these programs are also included electronically.

E. SAS Programs configured as SSGGmakeparfiles.sas. For each grade-subject, this program sorts the parameter data by test form, a configuration required by the FLUX program.

F. PARDUX Parameter Estimation Summary (SSGG04_SUM.TXT). This file provides a summary of the parameter estimation procedure run in PARDUX. It includes the limit data from the control file and also contains the number of stages PARDUX ran in order to reach convergence. It also contains the item numbers of items that could not be estimated and documents any items whose estimation reaches the maximum alpha parameter. This file identifies any problem items that might require additional manipulation before continuing the process.

G. PARDUX Parameter Estimation Details (SSGG04_DET.TXT). This file lists a systematic iteration of data, by item, during each stage of parameter estimation.

H. PARDUX Parameter File (SSGG04.PAR). This file contains parameter estimates for all items designated in the *.CTL file. It is used for later data manipulation.

I. PARDUX Item Summaries Files, Status (SSGG04_STAT.TXT). This file lists all items for a given test and their status after parameter estimation. Items are coded as either "estimate OK," "OK—default C," "not estimated," or "other codes." It provides a different type of record for the parameter estimation.

J. PARDUX Item Summaries Files, Distribution (SSGG04_DIST.TXT). This file contains the distribution of students who scored at each level on the open-response items. It is useful for examining the way that scoring rubrics for these items operate and for ensuring that all open-response items have the correct number of functioning score levels.

K. PARDUX Item Summaries Files, Parameters (SSGG04_PAR.TXT). This file contains the item parameters in different format from the *.PAR files. Word processing and spreadsheet programs can easily read this file.

L. PARDUX Item Summaries Files, Standard Errors (SSGG04_SE.TXT). This file contains the standard errors of estimation for each item including the errors for the various score levels on the open-response items.

M. PARDUX Item Summaries Files, FitQ1 (SSGG04_Q1.TXT). This file contains fit statistics for all items.

N. PARDUX Log File (SSGG04_LOG.TXT). As each manipulation of data is completed, PARDUX maintains a log of the procedures and filenames. This log is saved in text format.

O. Stocking-Lord Plots (SSGG04_SLPLOTS.doc). For each grade/subject combination, the Stocking-Lord data transformation calculates M1 and M2 values (slope and intercept) and outputs four graphs (one each for the a, b, and c parameters, and item p-values). The M1/M2 values, a log of the Stocking-Lord procedures, and the graphs are saved in this file.

P. FLUX control file (SSGG04.HLK). This file specifies the range of the scale scores as well as the M1 and M2 transformation constants from the Stocking-Lord transformation.

Q. FLUX Parameter Files by Form (SSGG041A.PAR, SSGG041B.PAR, etc.). Each parameter file computed using PARDUX was divided to represent items from each test form. Typically, 30 items were scored from each form. The exceptions are forms from Arts and Humanities and Practical Living/Vocational Studies, which each contain only 10 scored items.

R. Raw-Score-to-Scale-Score Tables (SSGG04_Flux.txt). A raw-score-to-scale-score table was produced for each form. These tables were saved in text format using FLUX.

S. Miscellaneous files and programs may also be included in the documentation. These files were constructed either during investigation of results or for future purposes.

Conclusion

CTB and HumRRO independently calculated the scaled/equated raw-score-to-scale-score tables for the 2004 Kentucky Core Content Test. From these tables, both identified cut points that could be used for assigning student performance classifications and later converted to school accountability indexes. No significant differences were found between CTB's and HumRRO's parameter estimation, Stocking-Lord transformation constants, raw-score-to-scale-score tables, or application of cut points. The differences that were found were in rounding of anchor item parameters – these rounding differences were so small that they had negligible effect on M1/M2 values and no effect on final cut points.

Additional scaled/equated raw-score-to-scale-score tables were calculated for reading and mathematics tests using only multiple-choice items. This one-time-only step was used to determine schools' preliminary AYP status. These calculations were made to meet reporting deadlines established by the federal NCLB Act. Differences between CTB and HumRRO were similarly small for the multiple-choice-only scale and had no effect on final cut points.

Given that the HumRRO and CTB scaling and linking results were nearly identical, HumRRO is confident that CTB did not commit processing errors.

References

Hoffman, R. G. & Thacker, A. A. (1999). *Third-party checking of 1999 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report SP-WATSD-99-44). Alexandria, VA: Human Resources Research Organization.

Thacker A. A. (2004). *Third-party checking of the creation of a multiple-choice-only scale for the Kentucky Core Content Test: for interim NCLB determinations in 2004.* (HumRRO Report FR-04-39). Alexandria, VA: Human Resources Research Organization.